数据中心网络中基于 SDN 的大象流负载均衡的研究 *

金　玲，束永安

(安徽大学 计算机科学与技术学院，合肥 230601)

摘　要：针对数据中心网络中大象流携带大量数据造成网络拥塞和负载不均衡的问题，提出基于 SDN(software defined network)的大象流负载均衡(elephant flow load balancing ，EFLB)。当网络负载超过阈值时，控制器利用 Openflow 特性将检测到的大象流分裂为多个老鼠流，并根据收集的网络拓扑和链路状态动态地计算负载最小的下一跳交换机，确保负载均衡。实验结果表明，相比于等价多路径路算法(equal-cost mulit-path routing ，ECMP)，EFLB 机制提高网络吞吐量和链路利用率，更好地实现网络负载均衡。

关键词：SDN；负载均衡；大象流；老鼠流

中图分类号：TP393.07　　doi: 10.3969/j.issn.1001-3695.2017.06.0655

Research on load balancing of elephant flow based on SDN In Data Center Network

Jin Ling[1], Shu Yongan[2]

(*College of Computer Science & Technology Anhui University*, *Hefei* 230601, *China*)

Abstract: To solve the problems of network congestion and load imbalance caused by elephant flow that carries large amounts of data in data center networks. This paper proposed an SDN based load balancing mechanism of elephant flow(EFLB) . When network load exceeded the threshold, the controller split the detected elephant flows down into multiple mice flows by Openflow feature and calculated dynamically the minimum load switch according to the collected network topology and link states to ensure load balancing. Experimental results show that the EFLB achieves network load balancing by improving the network throughput and link utilization compared with ECMP((equal-cost mulit-path routing) .

Key Words: SDN; load balancing; elephant flow; mice flow

0　引言

随着大数据和云计算的发展，网络的规模不断地扩大，会出现拥塞问题、延迟时间长和低吞吐量问题。软件定义网络(SDN)是将底层物理设备的控制功能分离出的新型范例。SDN 实现了网络流量的灵活控制和优化资源管理。控制器通过 OpenFlow[1]南向协议对交换机中流表进行操作，实现对数据平面的控制功能。OpenFlow 为每个流提供了三种统计计数(包、字节和持续时间)[2]。

Benson 等人[3]发现 80%的流小于 10KB(老鼠流)，而 10%的流则具有大量数据或生存周期长(大象流)。网络中大象流数量少于网络流数量的 10%，但却占网络流量的 80%[4]。研究发现在出口端的队列中老鼠流经常在大象流后面，造成长时间队列延迟[5,6]。为了高效管理网络，控制器是没有必要处理所有的流，只集中于对网络性能有重要影响的流的操作。因此，对大象流的识别和指定合适的转发策略是十分重要的。

为了解决数据中心网络由大象流造成的网络拥塞和负载不均衡问题，本文提出数据中心网络基于 SDN 大象流负载均衡机制(EFLB)。该机制以轮询方式监听网络，当网络负载超过阈值时，控制器将检测到的大象流分裂为多个老鼠流，并利用从数据平面收集的网络拓扑和链路状态动态计算出多个下一跳交换机中负载最小的交换机，用于发送分裂的老鼠流。该文使用 mininet 和胖树网络拓扑来评估提出的机制的性能。模拟实验结果证明，提出的 EFLB 机制明显提高网络吞吐量和链路利用率。

1　相关工作

目前，sFlow 是高速网络中主要的通过抽样检测大流的流量检测。然而检测大象流的准确性受到抽样比例的影响。Sun 等人[7]对检测大象流的各种数据包抽样方法进行比较，发现依赖抽样比独立抽样的准确性高。Lan 等人[8]提出 DLPO 算法，适用于不同的数据中心网络拓扑。通过在流传输过程中变换流路径，对网络流量进行负载均衡，提高了网络性能。

严军荣等人[9]提出一种大象流两级识别方法。提出一种大象流两级识别方法，在第一 阶段采用基于 TCP 发送队列的识

基金项目：安徽省自然科学基金资助项目（1408085MF125）

作者简介：金玲（1993-），女，安徽蚌埠人，硕士研究生，主要研究方向为软件定义网络（SDN）（826169566@qq.com）；束永安，（1966-），男，安徽合肥人，教授，博士，主要研究方向为计算机网络、无线网络.

别算法检测可疑大象流，在第二阶段采用基于流持续时间的识别算法检测真实大象流,减少流识别开销。Bi 等人[10]提出在 SDN 数据中心网络的二级自适应大象流检测系统，通过最小化正误率和负错误率，设计对检测系统的自适应性阈值，提高了对大象流识别的准确性，并减少了控制器处理大象流的负载。Liu 等人[11]提出基于 SDN 的大象流负载均衡机制(SD-LB)，提出根据动态计算多路径的权重的加权多路径路由算法。根据路径权重改变大象流的路径，从而减少网络负载。然而该方法只是对大象流更改路径，会造成目标路径上老鼠流的延迟。Xing 等人[12]提出基于数据包的采样与提取的二级大流检测机制，首先会检测出可疑的大象流，再使用基于流计数方法的 TCAM 来决定真实的大流，从而提高检测大象流的准确性。然而 TCAM 资源有限，会消耗 TCAM 资源。Chakraborty 等人[13]提出无须大象流检测的、并将大象流破坏成老鼠流的高效多路径路由机制，从而减少流完成时间。然而由于需要验证流是否是大象流，会增加成

本，而且会需要大量 TCAM 资源。Wang 等人[14]提出为大象流和老鼠流动态分配路径算法。根据全局网络状态，分别为大象流分配高吞吐量路径和为老鼠流规划低延迟路径，从而不仅保证老鼠流的低延迟而且保证了大象流的高吞吐量。Zhang 等人[15]提出通过找到大象流和核心/聚集交换机的稳定的匹配来提高大流的性能，从而避免拥塞。

2　EFLB 机制

针对数据中心网络的大象流造成的拥塞问题，本文提出基于 SDN 的大象流负载均衡机制(EFLB)，设置 EFLB 机制启动阈值 δ，控制器中的监控模块会实时监听数据平台，一旦发现平均吞吐量大于阈值 δ，就会启动 EFLB 机制平衡网络负载，防止网络阻塞。

首先终端主机检测大象流，将大象流分裂成老鼠流后，控制器根据动态计算的链路状态将老鼠流发送到多条路径，从而确保网络负载均衡。本文的 EFLB 机制框架如图 1 所示。

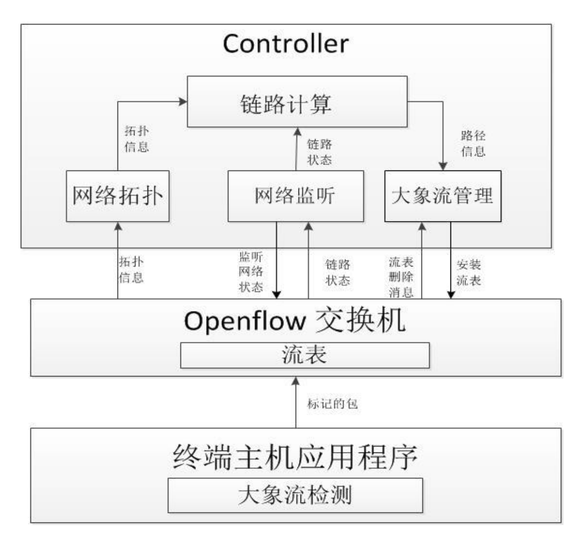


图 1　EFLB 机制的框架

2.1　大象流检测机制

EFLB 机制由于是将大象流分裂成老鼠流，再将其发送到多条路径，因此首先需要大象流检测机制。本文采用 Bi 等人[10]提出的在 SDN 数据中心网络的二级自适应大象流检测系统对大象流进行检测。它是在终端主机的应用程序对大象流进行检测，提高了对大象流检测的准确性。采用该检测系统检测出大象流，并对其进行标记。该检测系统首先采用包抽样来区分老鼠流和怀疑的大象流。将怀疑的大象流送到第二阶段，通过最小化正误率和负错误动态调整其阈值，从而提高检测大象流的准确性。

2.2　网络拓扑

这个模块处理、生成和分析网络中所有交换机的 LLDP 数据包。交换机将拓扑信息传给控制器，控制器中该模块将网络拓扑图传给链路计算模块用于链路的计算选择，并根据拓扑信息实时更新。

2.3　网络监听

首先，设置 EFLB 机制启动阈值 δ。如果网络负载大于该阈值 δ，EFLB 机制则会启动，检测出大象流，将其分裂为多个老鼠流，并根据网络链路状态将其发送到多个路径，防止网络拥塞。阈值 δ 的值在后面的网络胖树拓扑图实验中决定。

在该模块中，需要更新和存储所有 OpenFlow 交换机统计数据。这些统计数据需要用于链路计算模块来计算链路的负载。OpenFlow 协议提供了查询交换机的功能，以查找匹配特定规则或通过特定端口的流中的数据包数或字节数。对于网络上的每个链接，以字节数来收集实际的链接使用，使用这些信息来计算每个链路的实时可用带宽。此信息将帮助部署自定义路由协议，达到网络负载平衡和服务质量的目的。

为了获得统计数据，该模块需要在固定的间隔时间轮询 OpenFlow 交换机流量统计信息。支持 OpenFlow 的交换机可以提供每个流量统计信息，每个队列统计信息以及每个端口聚合统计信息。

获取整个流统计的一种简单的方法是固定间隔时间沿着每个流的路径轮询查询交换机，并合并其结果。然而，该策略对网络施加太多的通信开销，因为它以流为基础收集统计信息：重复请求和回复消息[16]。监控的准确性和网络开销取决于轮询交换机的频率。因此，有必要对流量行为进行自适应监控，当流量到达或改变其使用特性时，增加的轮询间隔，以及当流量统计收敛到稳定的行为时，减少轮询间隔。

测量的准确性还取决于可用计数器的数量和控制回路延迟：计数器的数量受到交换机和控制网络带宽上昂贵的功耗较大的 TCAM 数量的限制。控制回路延迟是从交换机发送计数器到控制器到安装新的计数规则的结束的延迟，包括传播延迟，控制器算法延迟和切换规则安装延迟[17]。对于较远的控制器实时收集测量信息和分析较为困难。因为发送测量数据的延迟和带宽的开销。

由于控制器需要从交换机收集实时流量统计信息，控制器

通过周期地向数据平面发送 Read-State 消息维持计数器和网络状态。当控制器请求流统计时，交换机应返回所有查询流的计数器。 这必然会在交换机上造成开销。当查询更多流条目时，CPU 负载增加。

“轮询全部交换机”策略被过度使用，则会由于从不同交换机重复收集相同的流量统计信息而带来额外的开销。因为多个交换机可能具有相同的流条目[16]。

2.4　大象流管理

将大象流分裂老鼠的基本思想是利用基于 Openflow 协议的流条目删除特征的 hard_time 值。根据 Openflow 协议的流表删除特征，流条目在交换机中持续的时间为 T，即 hard_time 属性的值为 T。该文在交换机安装一个 hard_time 值为 T 的主动流条目和 hard_time 值为 2T 备用流条目。当主动流条目经过时间 T 后被删除时，备用流条目变成主动流条目，并且将流切换到另一个链路，即备用流条目的出口端发生改变。因为大象流生存时间长，数据量大，大象流的出口再不断变化。最终，大象流被分裂成多个老鼠流。

当 EFLB 机制检测到大象流时，大象流管理模块需要安装一个主动流条目和备用流条目。因为流表的满期机制，第一个主动流条目在交换机中持续 T 时间后删除，交换机必须发送流条目删除消息给控制器。控制器接收到流条目删除消息，大象流管理模块必须安装 hard_time 值为 2T 并且优先级小于主动流条目的备用条目。而原先的备用流条目则变成主动流条目。

2.5　链路计算

为了防止数据中心网络大象流造成的网络阻塞，充分利用网络带宽。当检测到大象流时，利用 OpenFlow 的流条目删除特性将大象流分裂成老鼠流，链路计算模块会根据从交换机收集到的统计信息将分裂的老鼠流发送到多条链路来平衡网络负载。该模块的主要部分是链路优化算法。链路优化算法是网络监听模块从数据平面收集的统计信息和网络拓扑来计算交换机的负载，选择多个下一跳交换机中负载最小的交换机用于发送大象流分裂的老鼠流。

2.6　链路优化算法

本文用 $\lambda_{i,k}$ 表示当前第 k 个交换机的在接口 i 的吞吐量。交换机的负载相对于其他交换机的负载用式（1）表示。

$$L_{i,k}=\frac{\sum_i \lambda_{i,k}}{\sum_k\sum_j \lambda_{j,k}} \quad (1)$$

链路优化算法如下

```
while(大象流分裂结束)
  foreach(下一跳交换机 k 接口 i)
```

$$L_{i,k}=\frac{\sum_i \lambda_{i,k}}{\sum_k\sum_j \lambda_{j,k}}$$

```
  end
  sₐ={k|min(Lᵢ,ₖ)}
  将大象流分裂的老鼠流发送到 sₐ
end
```

3　性能评估

3.1　模拟环境搭建和设置模拟参数

该文使用 Floodlight 控制器和 mininet 仿真器，并使用胖树拓扑如图 2 所示。

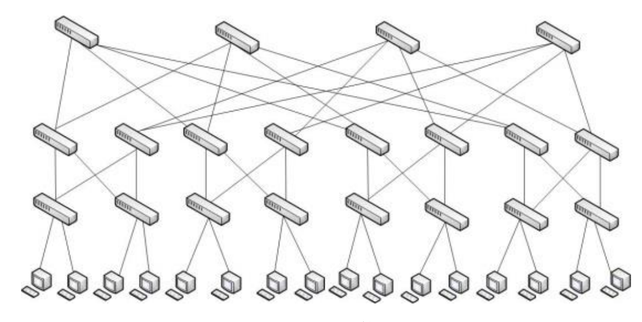


图 2　胖树拓扑 k=4

所有交换机都是 OpenFlow 交换机，控制平面使用 OpenFlow 1.3 与数据平面进行交互。在模拟实验之前，为了设置 EFLB 机制启动阈值 δ，本文计算在不同网络负载的平均吞吐量。如图 3 所示能够看到在网络负载在 800 M 时，吞吐量达到最大值，因此，本文选取吞吐量最大时网络负载的 70%为启动阈值 δ。

本文定义在 3 分钟之内携带流量超过全部网络流量的 0.03%的流为大象流并设置链路带宽和链路传播延迟分别为 100Mbps 和 1μs。在该实验中，用 Iperf 产生流量并遵从泊松分布。本文提出的 EFLB 机制与 ECMP 算法和 SD-LB 算法进行了比较。

3.2　实验结果

吞吐量和网络时延是评价网络性能的重要指标，反映了网络的拥塞程度。

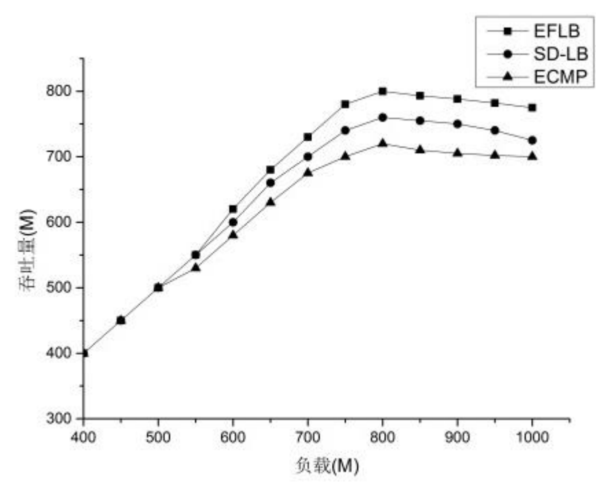


图 3　不同负载的网络吞吐量

如图 3 所示，发现 EFLB 机制明显提高了网络的吞吐量，这是由于 EFLB 机制将大象流分裂为多个老鼠流，并将其发送到负载小的路径。而 ECMP 算法没有改变大象流路径，当网络负载增大时，大象流带有大量数据会被阻塞。SD-LB 算法只是

将大象流更改路径。因为大象流携带大量数据，更改路径后可能会阻塞更改后的路径。

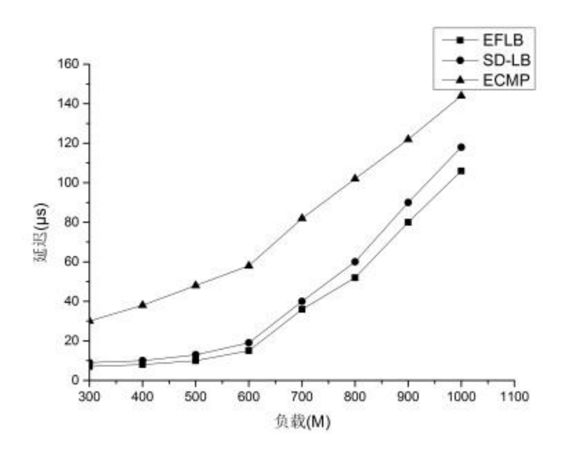


图 4　不同负载的网络时延

图 4 显示了随着网络负载的增大导致网络时延的变化。因为 ECMP 和 SD-LB 没有分裂大象流，老鼠流有可能在大象流后面会导致时延的增加。而 EFLB 将大象流分裂为老鼠流，并发送到多条链路，减小网络时延。

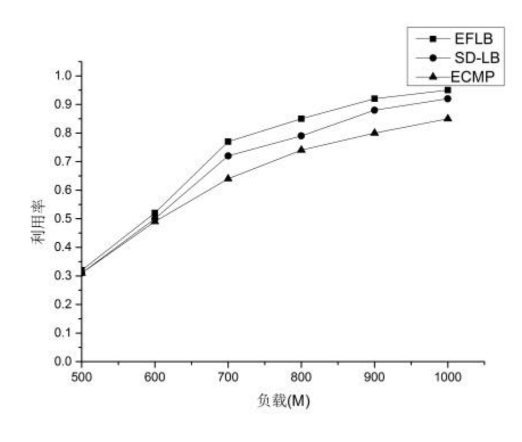


图 5 不同负载的链路利用率

图 5 显示了随着网络负载的增加导致链路利用率的变化。由于 ECMP 和 SD-LB 没有分裂大象流，只是改变了大象流的路径，并没有充分利用网络链路。而 EFLB 机制则分裂大象流为多个老鼠流，并发送到多条路径，明显提高了链路利用率。

4　结束语

由于大象流是网络中携带大量数据或生存周期长的流，会发现老鼠流经常在大象流队列后面，严重影响网络性能。所以针对大象流，提出数据中心网络基于 SDN 的大象流负载均衡机制(EFLB),控制器将检测到的大象流分裂为多个老鼠流，并利用从数据平面收集的网络拓扑和链路状态动态计算出多个下一跳交换机中负载最小的交换机，用于发送老鼠流。对提出的 EFLB 机制进行仿真实验，结果显示提出的 EFLB 机制明显提高了网络的吞吐量和链路利用率。下一步，进行分析 hard_time 值的对网络的影响，找到合适的 hard_time 值，将大象流分裂合适数量的老鼠流。

参考文献：

[1] Mckeown N, Anderson T, Balakrishnan H, et al. OpenFlow: enabling innovation in campus networks [J]. ACM SIGCOMM Computer Communication Review, 2008, 38 (2): 69-74.

[2] Kumar A, Sung M, Wang J, et al. Data streaming algorithms for efficient and accurate estimation of flow size distribution [C]// Proc of Joint International conference on Measurement and Modeling of Computer Systems. 2004: 177-188.

[3] Kandula S, Sengupta S, Greenberg A, et al. The nature of data center traffic: measurements & analysis [C]// Proc of ACM SIGCOMM Conference on Internet Measurement. 2009: 202-208.

[4] Benson T, Akella A, Maltz D A. Network traffic characteristics of data centers in the wild [C]// Proc of the 10th ACM SIGCOMM Conference on Internet Measurement. 2010: 267-280.

[5] Alizadeh M, Greenberg A, Maltz D A, et al. Data center TCP (DCTCP) [J]. ACM SIGCOMM Computer Communication Review, 2010, 40 (4): 63-74.

[6] Feamster N, Rexford J, Zegura E. The road to SDN: an intellectual history of programmable networks [J]. ACM SIGCOMM Computer Communication Review, 2014, 44 (2): 87-98.

[7] Sun Y, Liu W, Liu Z, et al. Comparison of five packet-sampling-based methods for detecting elephant flows [C]// Proc of Trustcom//BigDataSE//ISPA. 2016: 2018-2023.

[8] Lan Y L, Wang K, Hsu Y H. Dynamic load-balanced path optimization in SDN-based data center networks [C]// Proc of the 10th International Symposium on Communication Systems, Networks and Digital Signal Processing. 2016: 1-6.

[9] 严军荣，叶景畅，潘鹏. 一种大象流两级识别方法 [J]. 电信科学, 2017, 33 (3): 36-43.

[10] Bi C, Luo X, Ye T, et al. On precision and scalability of elephant flow detection in data center with SDN [C]// Proc of Globecom Workshops. 2013: 1227-1232.

[11] Liu J, Li J, Shou G, et al. SDN based load balancing mechanism for elephant flow in data center networks [C]// Proc of International Symposium on Wireless Personal Multimedia Communications. 2014: 486-490.

[12] Xing C, Ding K, Hu C, et al. Sample and fetch-based large flow detection mechanism in software defined networks [J]. IEEE Communications Letters, 2016, 20 (9): 1764-1767.

[13] Chakraborty S, Chen C. A low-latency multipath routing without elephant flow detection for data centers [C]// Proc of the 17th IEEE International Conference on High Performance Switching and Routing. 2016: 49-54.

[14] Wang W, Sun Y, Zheng K, et al. Freeway: adaptively isolating the elephant and mice flows on different transmission paths [C]// Proc of the 22nd IEEE International Conference on Network Protocols. 2014: 362-367.

[15] Zhang Y, Cui L, Chu Q. Fincher: elephant flow scheduling based on stable

matching in data center networks [C]// Proc of the 34th International Performance Computing and Communications Conference. 2015: 1-2.

[16] Su Z, Wang T, Xia Y, et al. FlowCover: low-cost flow monitoring scheme in software defined networks [C]// Proc of IEEE Global Communications Conference. 2015: 1956-1961.

[17] Moshref M, Yu M, Govindan R. Resource//accuracy tradeoffs in software-defined measurement [C]// Proc of ACM SIGCOMM Workshop on Hot Topics in Software Defined NETWORKING. 2013: 73-78.